บทความวิจัย

แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้เทคนิคการเรียนรู้ของเครื่อง

สุดา ทิพย์ประเสริฐ* ธรา อั่งสกุล และ จิติมนต์ อั่งสกุล
สาขาวิชาเทคโนโลยีสารสนเทศ คณะบริหารธุรกิจ มหาวิทยาลัยเทคโนโลยีสุรนารี

* ผู้นิพนธ์ประสานงาน โทรศัพท์ 08 6876 9035 อีเมล: d6110437@g.sut.ac.th DOI: 10.14416/j.kmutnb.2024.06.003
รับเมื่อ 6 มิถุนายน 2564 แก้ไขเมื่อ 7 สิงหาคม 2565 ตอบรับเมื่อ 13 กันยายน 2565 เผยแพร่ออนไลน์ 10 มิถุนายน 2567

บทคัดย่อ

ภาวะโรคซึมเศร้าเป็นหนึ่งในสาเหตุหลักที่ก่อให้เกิดปัญหาการฆ่าตัวตาย โดยผู้ป่วยเป็นโรคซึมเศร้าส่วนใหญ่ไม่ทราบว่าตนเองเกิดภาวะซึมเศร้าและมักแสดงออกผ่านทางสื่อสังคมออนไลน์ผ่านข้อความหรือรูปภาพ เนื่องจากสื่อสังคมออนไลน์เป็นรูปแบบการสื่อสารผ่านช่องทางที่ไม่ต้องอาศัยการแสดงออกทางสีหน้า นอกจากนี้งานวิจัยที่มีอยู่เน้นไปที่การใช้ข้อความในเครือข่ายสังคมออนไลน์มาวิเคราะห์เพียงอย่างเดียว งานวิจัยนี้จึงนำเสนอแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้เทคนิคการเรียนรู้ของเครื่อง โดยใช้ข้อมูลจากการทำแบบประเมินภาวะซึมเศร้า 9 คำถาม และจากผู้ใช้งานทวิตเตอร์ จำนวน 405 คน ร่วมกับเทคนิคการเรียนรู้ของเครื่องประกอบด้วย เทคนิคนาอีฟเบย์ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน เทคนิคต้นไม้การตัดสินใจ เทคนิคเพอร์เซ็ปตรอนหลายชั้น และเทคนิคการสุ่มป่าไม้ ผลการทดลองพบว่าแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าด้วยเทคนิคการสุ่มป่าไม้ให้ค่าประสิทธิภาพโดยรวมสูงที่สุด คือ ร้อยละ 87.39 ในขณะที่ผลการเปรียบเทียบประสิทธิภาพของแบบจำลองจากการใช้ตัวแปรนำเข้าที่แตกต่างกันพบว่า การใช้คุณลักษณะจากทวิตเตอร์ทั้งประเภทข้อความและรูปภาพให้ค่าประสิทธิภาพโดยรวมสูงที่สุด นอกจากนี้ การหาคุณลักษณะที่สำคัญพบว่า คุณลักษณะที่สำคัญที่สุดในการวิเคราะห์ความเสี่ยงภาวะซึมเศร้า คือ ค่าเฉลี่ยคะแนนความรู้สึกซึ่งเป็นคุณลักษณะของทวิตเตอร์ประเภทข้อความ

คำสำคัญ: ภาวะซึมเศร้า การวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า การเรียนรู้ของเครื่อง ทวิตเตอร์

การอ้างอิงบทความ: สุดา ทิพย์ประเสริฐ, ธรา อั่งสกุล และ จิติมนต์ อั่งสกุล, "แบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้เทคนิคการเรียนรู้ของเครื่อง," *วารสารวิชาการพระจอมเกล้าพระนครเหนือ*, ปีที่ 34, ฉบับที่ 3, หน้า 1–12, เลขที่บทความ 243-196144, ก.ค.–ก.ย. 2567.

Research Article

A Depression Risk Analysis Model using Machine Learning Techniques

Suda Tipprasert*, Thara Angskun and Jitimon Angskun
School of Information Technology, Faculty of Business Administration, Suranaree University of Technology, Nakhon Ratchasima, Thailand

* Corresponding Author, Tel. 08 6876 9035, E–mail: d6110437@g.sut.ac.th DOI: 10.14416/j.kmutnb.2024.06.003
Received 6 June 2021; Revised 7 August 2022; Accepted 13 September 2022; Published online: 10 June 2024

Abstract

Depression is one of the major causes of suicide. Most people with depression are unaware that they have depression and often express themselves on social media through text or images because social media is a form of communication through channels that do not rely on facial expressions. Moreover, the existing research focuses solely on the use of texts in social networks for analysis. This research proposes a depression risk analysis model using machine learning techniques. The model construction applies data from the Patient Health Questionnaire-9 and Twitter data in combination with machine learning techniques. The Twitter data are collected from 405 Twitter users. There are five machine learning techniques explored in this research which are Naïve Bayes, Support Vector Machine, Decision Tree, Multilayer Perceptron, and Random Forest. The experimental results indicated that the depression risk analysis model using a Random Forest technique provided the highest F-measure at 87.39%. While a comparison of the model's performance using different input variables revealed that the use of Twitter attributes, both text and image, achieved the highest F-measure. In addition, the feature finding revealed that the most important feature in the depression risk analysis was the mean sentiment score, which is a Twitter's text attribute.

Keywords: Depression, Depression Risk Analysis, Machine Learning, Twitter

Please cite this article as: S. Tipprasert, T. Angskun and J. Angskun, "A Depression Risk Analysis Model using Machine Learning Techniques," *The Journal of KMUTNB*, vol. 34, no. 3, pp. 1–12, ID. 243-196144, Jul.–Sep. 2024 (in Thai).

1. บทนำ

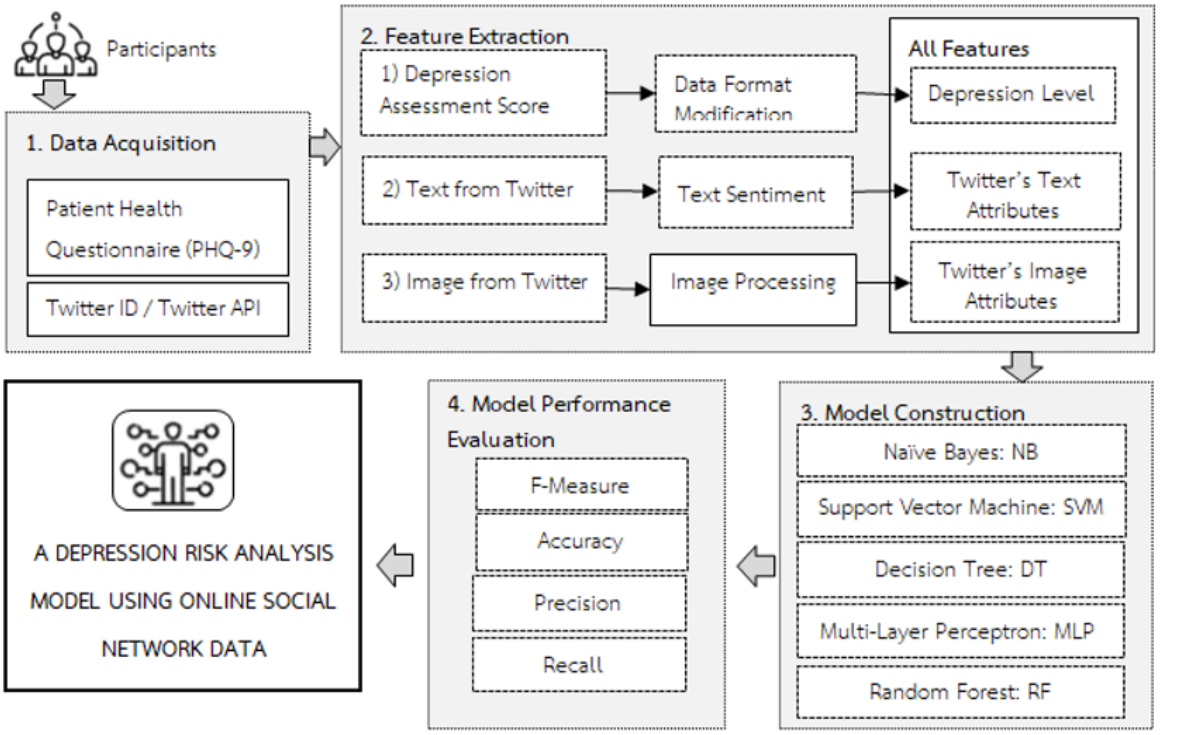
ภาวะโรคซึมเศร้าเป็นหนึ่งในสาเหตุหลักที่ก่อให้เกิดปัญหาการฆ่าตัวตายตามมา ใน พ.ศ. 2560 ผู้ที่ได้รับการค้นหาคัดกรองว่า มีแนวโน้มป่วยเป็นโรคซึมเศร้าด้วยแบบคัดกรอง 2Q มีจำนวนถึง 14 ล้านคน และมีถึงร้อยละ 64 ที่ไม่ได้รับการรักษา [1] อัตราการป่วยโรคซึมเศร้าในประเทศไทยได้เพิ่มขึ้นอย่างต่อเนื่องในช่วง พ.ศ. 2552–2562 โดยใน พ.ศ. 2562 พบผู้ป่วยโรคซึมเศร้าเพิ่มขึ้นถึงร้อยละ 25.92 [2] และใน พ.ศ. 2563 มีอัตราการฆ่าตัวตายสำเร็จประมาณ 4,000 ราย ซึ่งร้อยละ 60 มีสาเหตุมาจากโรคซึมเศร้า [3] แสดงให้เห็นถึงความจำเป็นและความสำคัญในการสนับสนุนงานด้านการป้องกัน รักษา และส่งเสริมสุขภาวะทางจิตของคนไทย [4]

จากการเพิ่มขึ้นของจำนวนผู้ป่วยเป็นโรคซึมเศร้า และส่วนใหญ่ไม่ได้รับการรักษา หรือไม่ทราบว่าตนเองเกิดภาวะซึมเศร้าจึงแสดงออกผ่านทางสื่อสังคมออนไลน์เนื่องจากสื่อสังคมออนไลน์เป็นรูปแบบการสื่อสารผ่านช่องทางที่ไม่ต้องอาศัยน้ำเสียง การสบตา และการแสดงออกทางสีหน้า [5] ซึ่งสามารถแสดงออกได้จากการแสดงความคิดเห็นผ่านข้อความ หรือรูปภาพ

จากการทบทวนงานวิจัยที่เกี่ยวข้องกับการวิเคราะห์ภาวะซึมเศร้าจากเครือข่ายสังคมออนไลน์ เช่น การจำแนกประเภทโดยใช้ปัญญาประดิษฐ์ในการพัฒนาแบบจำลองเพื่อทำนายระดับภาวะซึมเศร้า โดยใช้ข้อความแสดงความคิดเห็นจากเครือข่ายสังคมออนไลน์ด้วยขั้นตอนวิธีซัพพอร์ตเวกเตอร์แมชชีน และขั้นตอนวิธีนาอีฟเบย์ มีค่าความถูกต้องอยู่ที่ร้อยละ 57 และร้อยละ 63 ตามลำดับ [6] และมีงานวิจัยที่ใช้เทคนิคซัพพอร์ตเวกเตอร์แมชชีน ในการทำนายภาวะซึมเศร้าผ่านเครือข่ายสังคมออนไลน์ โดยการรวบรวมข้อมูลการใช้งานทวิตเตอร์จำนวน 476 คน โดยให้ค่าความถูกต้อง (Accuracy) ร้อยละ 70 และพบว่า ข้อความทวีตด้านลบมีสัญญาณที่บ่งบอกถึงภาวะซึมเศร้า [7] นอกจากนั้นมีการประยุกต์ใช้การวิเคราะห์ความรู้สึกร่วมกับการเรียนรู้เชิงลึก (Deep Learning) ในการจำแนกความรู้สึกของข้อมูลในทวิตเตอร์ ผลการวิจัยพบว่า เทคนิคการเรียนรู้เชิงลึกให้ผลลัพธ์ด้านความถูกต้องที่ดีกว่าเทคนิคดั้งเดิม เช่น นาอีฟย์เบย์ ซัพพอร์ตเวกเตอร์แมชชีน โดยมีค่าความถูกต้องอยู่ที่ร้อยละ 75 [8] และงานวิจัยที่นำเสนอผลลัพธ์การสร้างเฟรมเวิร์คในการวิเคราะห์ความรู้สึกโดยผู้ใช้งานทวิตเตอร์ โดยได้วิเคราะห์ข้อความคิดเห็นจากทวิตเตอร์ในระยะเวลา 2 เดือน ซึ่งข้อความในทวิตเตอร์ส่วนใหญ่มีข้อมูลเกี่ยวกับความรู้สึกซึมเศร้า สถานะ ประวัติการรักษา จากการศึกษาพบว่า การใช้ข้อความด้านลบ ข้อความที่แสดงออกถึงความซึมเศร้า หรือสัญรูปอารมณ์ด้านลบผ่านทางทวิตเตอร์ที่เพิ่มขึ้นนั้นมีนัยสำคัญกับการเกิดอาการโรคซึมเศร้า [9] นอกจากนั้นยังมีงานวิจัยที่พัฒนาแบบจำลองการทำนายภาวะโรคซึมเศร้าจากรูปภาพในอินสตราแกรม (Instagram) โดยใช้ขั้นตอนวิธีการวิเคราะห์การถดถอยโลจิสติแบบเบย์ (Bayesian Logistic Regression) กลุ่มตัวอย่างจำนวน 166 คน และใช้รูปภาพจำนวน 43,950 ภาพ ในการวิเคราะห์สี (Color Analysis) และการตรวจสอบใบหน้า (Face Detection) ผลการทดลองพบว่า แบบจำลองมีประสิทธิภาพดีเมื่อเทียบกับการวินิจฉัยแบบทั่วไป โดยมีค่าประสิทธิภาพโดยรวมอยู่ที่ร้อยละ 61 [10]

จากการศึกษางานวิจัยที่เกี่ยวข้องนั้นพบว่า งานวิจัยส่วนใหญ่เน้นไปที่การใช้ข้อความ หรือคุณลักษณะของข้อความในเครือข่ายสังคมออนไลน์ เช่น จำนวนข้อความทวีต จำนวนคำด้านบวก จำนวนคำด้านลบ ข้อความที่บ่งบอกถึงภาวะซึมเศร้า [7], [9] มาวิเคราะห์ ซึ่งการวิเคราะห์ภาวะซึมเศร้าจากข้อความแสดงความคิดเห็นทวิตเตอร์เพียงอย่างเดียวอาจไม่เพียงพอ และมีเพียงงานวิจัยเดียวที่ใช้รูปภาพในเครือข่ายสังคมออนไลน์มาวิเคราะห์ แต่อย่างไรก็ตามยังไม่มีการนำทั้งข้อความและรูปภาพมาวิเคราะห์ร่วมกัน นอกจากนั้นยังพบว่า งานวิจัยส่วนใหญ่มีค่าความถูกต้องอยู่ในระดับต่ำกว่าร้อยละ 80 อีกทั้งงานวิจัยบางส่วนไม่ได้มีการเปรียบเทียบเทคนิคอื่น ๆ เพื่อหาวิธีที่ดีที่สุด และยังไม่พบงานวิจัยใดที่วิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าที่สามารถบ่งบอกถึงระดับคะแนนภาวะซึมเศร้าได้

ดังนั้นงานวิจัยนี้จึงนำคุณลักษณะของทวิตเตอร์ที่เป็นทั้งข้อความและรูปภาพ เพื่อพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้เทคนิคการ



รูปที่ 1 กรอบการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์

เรียนรู้ของเครื่อง (Machine Learning) ได้แก่ เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine; SVM) เทคนิคนาอีฟเบย์ (Naïve Bayes) เทคนิคต้นไม้การตัดสินใจ (Decision Tree) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron) และเทคนิคแรนดอมฟอเรส (Random Forest) เพื่อเปรียบเทียบผลลัพธ์และวิเคราะห์หาคุณลักษณะและเทคนิคที่เหมาะสมที่สุดที่นำมาใช้ในการพัฒนาแบบจำลองนั้น

2. วัสดุ อุปกรณ์และวิธีการวิจัย

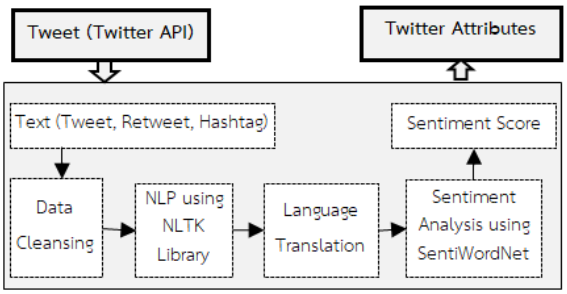
การพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ ประกอบด้วย 4 ขั้นตอน โดยมีกรอบการทำงานดังแสดงในรูปที่ 1

2.1 การรวบรวมข้อมูล (Data Acquisition)

ในการรวบรวมข้อมูลสำหรับการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะโรคซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ ใช้ข้อมูลจากการทำแบบประเมินภาวะซึมเศร้า 9 คำถาม (PHQ-9) [11] ร่วมกับข้อมูลการใช้งานทวิตเตอร์ (Twitter ID) จำนวน 405 คน แบ่งเป็น 4 กลุ่ม ได้แก่ ผู้ที่ไม่มีภาวะซึมเศร้า (Level0) จำนวน 154 คน มีภาวะซึมเศร้าระดับน้อย (Level1) จำนวน 149 คน มีภาวะซึมเศร้าระดับปานกลาง (Level2) จำนวน 70 คน และมีภาวะซึมเศร้าระดับรุนแรง (Level3) จำนวน 32 คน

2.2 การสกัดคุณลักษณะของข้อมูล (Feature Extraction)

ในขั้นตอนนี้เป็นการสกัดคุณลักษณะจากข้อมูล 3 ส่วน คือ 1) ข้อมูลระดับคะแนนภาวะซึมเศร้า (Depression Assessment Score) 2) ข้อมูลคุณลักษณะจากทวิตเตอร์ประเภทข้อความ (Text from Twitter) และ 3) ข้อมูลคุณลักษณะจากทวิตเตอร์ประเภทรูปภาพ (Image from Twitter) โดยมีรายละเอียดดังต่อไปนี้



รูปที่ 2 การสกัดคุณลักษณะจากข้อมูลประเภทข้อความในทวิตเตอร์

2.2.1 การสกัดระดับคะแนนภาวะซึมเศร้า

ระดับคะแนนภาวะซึมเศร้าเป็นข้อมูลที่ได้จากการทำแบบประเมินภาวะซึมเศร้า 9 คำถาม โดยเปลี่ยนรูปแบบข้อมูล (Data Format Modification) จากข้อมูลเชิงตัวเลข (Numeric) เป็นข้อมูลนามบัญญัติ (Nominal) ซึ่งใช้เกณฑ์การเปลี่ยนข้อมูลตามแบบสอบถามภาวะซึมเศร้า 9 คำถาม คือ คะแนนรวมน้อยกว่า 7 ไม่มีอาการของภาวะซึมเศร้า คะแนนรวม 7–12 มีภาวะซึมเศร้าระดับน้อย คะแนนรวม 13–18 มีภาวะซึมเศร้าระดับปานกลาง คะแนนรวมมากกว่า 19 คะแนน มีภาวะซึมเศร้าระดับรุนแรง

2.2.2 การสกัดคุณลักษณะจากทวิตเตอร์ประเภทข้อความ

ในขั้นตอนนี้เป็นการนำข้อความจากทวิตเตอร์ (Text) ได้แก่ ทวีต (Tweet) รีทวีต (Retweet) และแฮชแท็ก (Hashtag) เข้าสู่กระบวนการทำความสะอาดข้อมูล (Data Cleansing) เพื่อตัดข้อความหรือสัญลักษณะที่ไม่เกี่ยวข้องออก หลังจากนั้นจึงเข้าสู่กระบวนการประมวลผลภาษาธรรมชาติ (Natural Language Processing; NLP) โดยใช้ไลบารี NKTK ในการตัดพยางค์ย่อย (Tokenization) เพื่อแยกข้อความในรูปประโยค (Sentence) ออกเป็นคำ (Word) หรือเรียกว่าการตัดคำ (Word Segmentation) โดยใช้ขั้นตอนวิธีการตัดคำแบบเหมือนมากที่สุด (Maximum Matching Algorithm) และเข้าสู่กระบวนการแปลภาษา (Language Translation) โดยแปลจากภาษาไทยเป็นภาษาอังกฤษ ด้วย translate 3.6.1 Python Library ใน PyPi (The Python Package Index) ซึ่งเป็นแหล่งรวมชุดคำสั่งของ Python และเป็นเครื่องมือแปลภาษาที่ใช้งานง่าย รองรับผู้ให้บริการที่หลากหลาย ซึ่งในงานวิจัยนี้ใช้ผู้ให้บริการในการแปลภาษา คือ MyMemory API จากนั้นนำผลลัพธ์ที่ได้มาเข้าสู่กระบวนการวิเคราะห์ความรู้สึกด้วยฐานข้อมูลคำศัพท์ (SentiWordNet) ซึ่งเป็นคลังคำศัพท์ที่ใช้ในการวิเคราะห์ความรู้สึกที่ได้รับความนิยมสูงและมีคำศัพท์มากที่สุด โดยมีมากถึง 117,695 คำ [12]–[15] โดยนำผลลัพธ์ที่ได้มาค้นหาคำศัพท์ในฐานข้อมูลคำศัพท์ที่บอกระดับคำเป็นบวก และเป็นลบ เพื่อคำนวณหาค่าคะแนนความรู้สึก (Sentiment Score) ดังแสดงในรูปที่ 2

โดยแบ่งเป็น คะแนนด้านบวก ($S+$) และ คะแนนด้านลบ ($S-$) ได้ดังสมการที่ (1) และ (2) ตามลำดับ โดยที่ n คือ จำนวนคำทั้งหมด

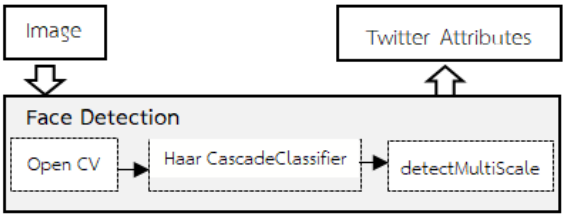
$$S+ = \frac{\sum_{i \in t} PosScore_i}{n} \quad (1)$$

$$S- = \frac{\sum_{i \in t} NegScore_i}{n} \quad (2)$$

จากนั้นสามารถคำนวณได้ว่าประโยคที่ได้นั้นเป็นประโยคด้านบวก (Positive) หรือด้านลบ (Negative) จากเงื่อนไขดังนี้

```
if  S+ > S– :
        Sentence is positive
if  S+ <= S– :
        Sentence is negative
```

ในส่วนของจำนวนทวีตที่แสดงถึงภาวะซึมเศร้าได้จาก ผลรวมของทวีตที่มีคำที่แสดงถึงภาวะซึมเศร้า (Depression Term Corpus) [16] โดยการคำนวณคะแนนจะทำให้สามารถจำแนกคุณลักษณะจากทวิตเตอร์ประเภทข้อความได้ 10 คุณลักษณะ ได้แก่ จำนวนทวีตด้านบวก 429 ทวีต จำนวนทวีตด้านลบ 570 ทวีต จำนวนทวีตที่แสดงถึงภาวะซึมเศร้า 100 ทวีต จำนวนรีทวีตด้านบวก



รูปที่ 3 การสกัดคุณลักษณะจำนวนบุคคลในรูปภาพจากการตรวจจับใบหน้า

3,490 ทวีต จำนวนรีทวีตด้านลบ 4,158 ทวีต จำนวนรีทวีตที่แสดงถึงภาวะซึมเศร้า 670 ทวีต จำนวนแฮชแท็กด้านบวก 52 แฮชแท็ก จำนวนแฮชแท็กด้านลบ 58 แฮชแท็ก จำนวนแฮชแท็กที่แสดงถึงภาวะซึมเศร้า 27 แฮชแท็ก และค่าเฉลี่ยคะแนนความรู้สึกของทวีต

2.2.3 การสกัดคุณลักษณะจากทวิตเตอร์ประเภทรูปภาพ

การสกัดคุณลักษณะจากรูปภาพ (Image) ที่ได้จากการโพสต์ในทวิตเตอร์ ทำโดยการนับจำนวนบุคคลในรูปภาพและการวิเคราะห์สีของรูปภาพ

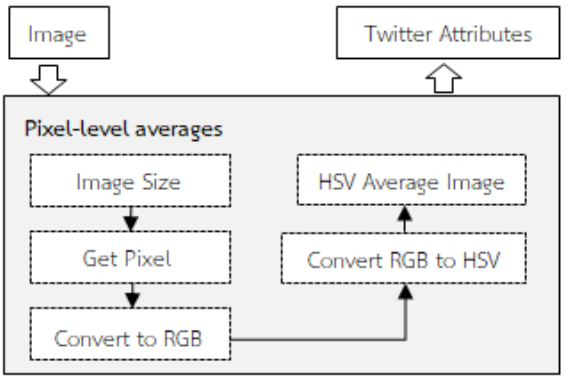
การนับจำนวนบุคคลในภาพใช้ซอฟต์แวร์การตรวจจับใบหน้า (Face Detection) โดยใช้ตัวแยกประเภทการเรียงซ้อนตามฟีเชอร์ของฮาร์ (Haar Feature-based Cascade Classifiers) เป็นวิธีการตรวจจับวัตถุที่มีประสิทธิภาพซึ่งเป็นส่วนหนึ่งของไลบรารี OpenCV [17] ซึ่งขั้นตอนการสกัดคุณลักษณะจำนวนบุคคลในรูปภาพ แสดงได้ดังรูปที่ 3

การวิเคราะห์ค่าสีของรูปภาพใช้การวิเคราะห์ค่าเฉลี่ยระดับพิกเซล (Pixel-level Averages) ที่ได้จากการคำนวณค่าสี 3 รูปแบบคือ เฉดสี (Hue; H) ความอิ่มตัวของสี (Saturation: S) และความสว่างของสี (Value; V) ดังแสดงในรูปที่ 4

เมื่อได้ค่า HSV ในระดับพิกเซลแล้ว จึงนำไปคำนวณหาค่าสีเฉลี่ยของรูปภาพจากสมการที่ (3)–(5)

$$Pixel\ AVG\ Hue = \frac{\sum Hue}{ImageSize} \quad (3)$$

$$Pixel\ AVG\ Saturation = \frac{\sum Saturation}{ImageSize} \quad (4)$$



รูปที่ 4 ขั้นตอนการวิเคราะห์ค่าสีของรูปภาพโดยใช้การวิเคราะห์ค่าเฉลี่ยระดับพิกเซล

$$Pixel\ AVG\ Value = \frac{\sum Value}{ImageSize} \quad (5)$$

หลังจากนั้นจึงนำค่าเฉลี่ยความอิ่มตัวของสี (Pixel AVG Saturation) และค่าเฉลี่ยความสว่างของสี (Pixel AVG Value) ไปใช้สำหรับวิเคราะห์ว่าเป็นรูปภาพสีสดใส หรือรูปภาพสีหม่น โดยมีเงื่อนไข [10], [18] ดังนี้

```
If Pixel AVG Value >=50 and Pixel AVG Saturation >=50 :
    Image is Colorful image
Else :
    Image is Dull image
```

หากค่าเฉลี่ยความอิ่มตัวของสี (Pixel AVG Saturation) และค่าเฉลี่ยความสว่างของสี (Pixel AVG Value) มีค่ามากกว่าหรือเท่ากับ 50 แสดงว่าภาพนั้นเป็นภาพสีสดใส (Colorful Image) ถ้าค่าเฉลี่ยความอิ่มตัวของสี หรือค่าเฉลี่ยความสว่างของสี มีค่าน้อยกว่า 50 แสดงว่าภาพนั้นเป็นภาพสีหม่น (Dull Image)

โดยค่าของข้อมูลในแต่ละคุณลักษณะจากทวิตเตอร์ประเภทรูปภาพ ประกอบด้วย จำนวนรูปภาพสีสดใส 3,660 รูป จำนวนรูปภาพสีหม่น 1,471 รูป จำนวนรูปภาพที่ไม่มีคน 2,154 รูป จำนวนรูปภาพที่มีคน 1 คน 1,452 รูป

และจำนวนรูปภาพที่มีคนมากกว่า 1 คน 1,422 รูป

ในขั้นตอนการสกัดคุณลักษณะจากข้อมูลทั้ง 3 ส่วน แสดงคุณลักษณะทั้งหมด (All Features) ได้ดังตารางที่ 1

ตารางที่ 1 คุณลักษณะทั้งหมด (All Features)

รหัส	คุณลักษณะ	ชนิดข้อมูล	รายละเอียด
คุณลักษณะจากทวิตเตอร์ประเภทข้อความ (Text)			
X_1	Positive Tweets	Numeric	จำนวนทวีตด้านบวก
X_2	Negative Tweets	Numeric	จำนวนทวีตด้านลบ
X_3	Depression Tweets	Numeric	จำนวนทวีตที่แสดงถึงภาวะซึมเศร้า
X_4	Positive ReTweets	Numeric	จำนวนรีทวีตด้านบวก
X_5	Negative ReTweets	Numeric	จำนวนรีทวีตด้านลบ
X_6	Depression ReTweets	Numeric	จำนวนรีทวีตที่แสดงถึงภาวะซึมเศร้า
X_7	Positive Hashtags	Numeric	จำนวนแฮชแท็กด้านบวก
X_8	Negative Hashtags	Numeric	จำนวนแฮชแท็กด้านลบ
X_9	Depression Hashtags	Numeric	จำนวนแฮชแท็กที่แสดงถึงภาวะซึมเศร้า
X_{10}	Sentiment Score	Numeric	ค่าเฉลี่ยคะแนนความรู้สึกของทวีต
คุณลักษณะจากทวิตเตอร์ประเภทรูปภาพ (Image)			
X_{11}	Colorful Image	Numeric	จำนวนรูปภาพสีสดใส
X_{12}	DullImage	Numeric	จำนวนรูปภาพสีหม่น
X_{13}	NoPerson-Image	Numeric	จำนวนรูปภาพที่ไม่มีคน
X_{14}	OnePerson Image	Numeric	จำนวนรูปภาพที่มีคน 1 คน
X_{15}	ManyPerson Image	Numeric	จำนวนรูปภาพที่มีคนมากกว่า 1 คน

ตารางที่ 1 คุณลักษณะทั้งหมด (All Features) (ต่อ)

รหัส	คุณลักษณะ	ชนิดข้อมูล	รายละเอียด
Y	depression	Nominal	ภาวะซึมเศร้า Level 0: ไม่มีภาวะซึมเศร้า Level 1: ภาวะซึมเศร้าระดับน้อย Level 2: ภาวะซึมเศร้าระดับปานกลาง Level 3: ภาวะซึมเศร้าระดับรุนแรง

2.3 การสร้างแบบจำลองที่ใช้ในการพยากรณ์ (Model Construction)

ในขั้นตอนนี้เป็นการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่อง โดยใช้ข้อมูลการใช้งานทวิตเตอร์ของผู้ใช้ โดยมีตัวแปรนำเข้า (Input Variables) คือ $X_1 - X_{15}$ และตัวแปรค่าเป้าหมาย (Target Variable) คือ Y ด้วยวิธีการไขว้ทบ 10 ส่วน (10-fold cross validation) เพื่อประเมินแบบจำลอง ซึ่งการทดลองนี้ได้ใช้ Python Library สำหรับการประมวลผลการทดลอง และใช้ Scikit-learn Library ร่วมกับ Keras library และ Tensorflow Library สำหรับการสร้างแบบจำลองด้วยเทคนิคการเรียนรู้ของเครื่อง 5 เทคนิค ได้แก่ 1) นาอีฟเบย์ (Naïve Bayes; NB) 2) ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) โดยใช้เคอร์เนลฟังก์ชัน (Kernel Function) คือเรเดียลเบสิสฟังก์ชัน (Radial Basis Function; RBF) 3) ต้นไม้การตัดสินใจ (Decision Tree; DT) โดยใช้ขั้นตอนวิธี C4.5 4) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron; MLP) โดยมีชั้นซ่อน (Hidden Layers) คือ 100 และฟังก์ชั้นกระตุ้น (Activation Function) คือ ReLu และ 5) เทคนิคการสุ่มป่าไม้ (Random Forest; RF) โดยใช้จำนวนต้นไม้ (Tree) คือ 100 ซึ่งเทคนิคการเรียนรู้ของเครื่องทั้ง 5 เทคนิค เป็นเทคนิคการเรียนรู้ของเครื่องแบบมีผู้สอน (Supervised Machine Learning) และเป็นเทคนิคที่ใช้ในการจำแนกหมวดหมู่ (Classification) ที่ได้รับความนิยมในการพัฒนาแบบจำลองในการตรวจจับภาวะซึมเศร้าจากข้อมูลเครือข่ายสังคมออนไลน์ [5]–[8], [16]

2.4 การประเมินประสิทธิภาพแบบจำลอง

การประเมินประสิทธิภาพแบบจำลองใช้การวัดค่ามาตรฐาน 4 ค่า ได้แก่ ค่าความถูกต้อง (Accuracy) ค่าความแม่นยำ (Precision) ค่าความระลึก (Recall) และค่าประสิทธิภาพโดยรวม (F-Measure) ดังสมการที่ (6)–(9)

$$Accuracy_c = \frac{TP_c + TN_c}{(TP_c + TN_c + FP_c + FN_c)} \times 100$$

$$Accuracy = \frac{\sum_{c=0}^{3} Accuracy_c}{4} \quad (6)$$

$$Precision_c = \frac{TP_c}{(TP_c + FP_c)} \times 100$$

$$Precision = \frac{\sum_{c=0}^{3} Precision_c}{4} \quad (7)$$

$$Recall_c = \frac{TP_c}{(TP_c + FN_c)} \times 100$$

$$Recall = \frac{\sum_{c=0}^{3} Recall_c}{4} \quad (8)$$

$$F1_c = \frac{2 \times (Recall_c \times Precision_c)}{(Recall_c + Precision_c)}$$

$$F - Measure = \frac{\sum_{c=0}^{3} F1_c}{4} \quad (9)$$

โดยที่

TP_c (True Positive) คือ จำนวนกลุ่มตัวอย่างที่อยู่ในกลุ่ม c และแบบจำลองทำนายว่าอยู่ในกลุ่ม c

TN_c (True Negative) คือ จำนวนกลุ่มตัวอย่างที่ไม่อยู่ในกลุ่ม c และแบบจำลองทำนายว่าไม่อยู่ในกลุ่ม c

FP_c (False Positive) คือ จำนวนกลุ่มตัวอย่างที่ไม่อยู่ในกลุ่ม c แต่แบบจำลองทำนายว่าอยู่ในกลุ่ม c

FN_c (False Negative) คือ จำนวนกลุ่มตัวอย่างที่อยู่ในกลุ่ม c แต่แบบจำลองทำนายว่าไม่อยู่ในกลุ่ม c

c คือ กลุ่มของระดับคะแนนจากผลการประเมินภาวะซึมเศร้าที่ทดสอบด้วยแบบทดสอบภาวะซึมเศร้า 9 คำถาม (9 Questionnaires; 9Q) ซึ่งมี 4 ระดับคะแนน เมื่อ $0 \le c \le 3$

3. ผลการทดลอง

งานวิจัยนี้ได้แบ่งผลการวิจัยออกเป็น 3 ส่วน ดังนี้

3.1 ผลการประเมินประสิทธิภาพของแบบจำลอง

เมื่อพิจารณาผลการเปรียบเทียบค่าประสิทธิภาพของแบบจำลองเมื่อใช้คุณลักษณะทั้งหมดในการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้เทคนิคการเรียนรู้ของเครื่อง ผลที่ได้แสดงดังตารางที่ 2 พบว่า แบบจำลองที่พัฒนาด้วยเทคนิคการสุ่มป่าไม้ (RF) ให้ค่าประสิทธิภาพโดยรวมสูงที่สุด คือ ร้อยละ 87.39 และให้ค่าความถูกต้องสูงที่สุดคือร้อยละ 86.47 รองลงมา คือ เทคนิคต้นไม้การตัดสินใจ (DT) เทคนิคซัพพอร์ตเวกเตอร์แมชชีน (SVM) เทคนิคเพอร์เซ็ปตรอนหลายชั้น (MLP) และเทคนิคนาอีฟเบย์ (NB) ตามลำดับ

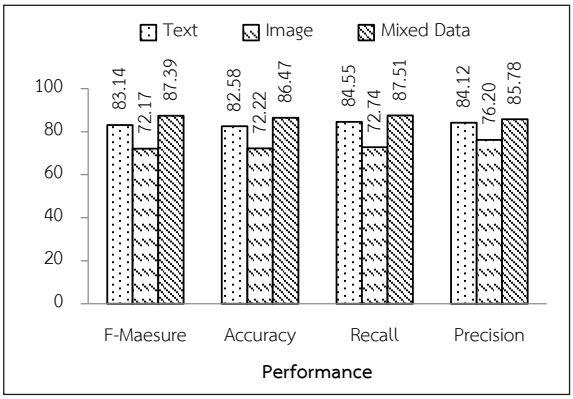
ตารางที่ 2 ผลการเปรียบเทียบค่าประสิทธิภาพแบบจำลอง

แบบจำลอง	ค่าประสิทธิภาพโดยรวม	ค่าความถูกต้อง	ค่าความแม่นยำ	ค่าความระลึก
SVM	80.52	81.71	84.17	80.76
DT	82.19	82.94	82.83	81.56
NB	26.43	36.06	47.01	33.92
RF	87.39	86.47	87.51	85.78
MLP	75.59	75.50	77.45	75.44

โดยที่ SVM คือ ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine; SVM)
DT คือ ต้นไม้การตัดสินใจ (Decision Tree; DT)
NB คือ นาอีฟเบย์ (Naïve Bayes; NB)
RF คือ การสุ่มป่าไม้ (Random Forest; RF)
MLP คือ เพอร์เซ็ปตรอนหลายชั้น (Multilayer Perceptron; MLP)

3.2 การเปรียบเทียบประสิทธิภาพของแบบจำลองจากการใช้ตัวแปรนำเข้าที่แตกต่างกัน

การเปรียบเทียบในขั้นตอนนี้ทำโดยการสร้างแบบจำลองด้วยเทคนิคการสุ่มป่า (RF) โดยใช้ตัวแปรนำเข้าที่แตกต่างกัน 2 ประเภท ได้แก่ คุณลักษณะจากทวิตเตอร์ประเภทข้อความ (Text from Twitter) (ตัวแปรนำเข้า



รูปที่ 5 การเปรียบเทียบประสิทธิภาพของแบบจำลองจากการใช้ตัวแปรนำเข้าที่แตกต่างกัน

$X_1 - X_{10}$) และคุณลักษณะจากทวิตเตอร์ประเภทรูปภาพ (Image from Twitter) (ตัวแปรนำเข้า $X_{11} - X_{15}$) ซึ่งผลการเปรียบเทียบประสิทธิภาพ แสดงได้ดังรูปที่ 5 จะเห็นได้ว่าคุณลักษณะจากทวิตเตอร์ประเภทรูปภาพให้ค่าประสิทธิภาพโดยรวมน้อยกว่าคุณลักษณะจากทวิตเตอร์ประเภทข้อความ ในขณะที่เมื่อนำคุณลักษณะทั้งสอง คือ คุณลักษณะจากทวิตเตอร์ประเภทข้อความ และประเภทรูปภาพมาใช้ร่วมกัน (Mixed Data) ให้ค่าประสิทธิภาพโดยรวมสูงที่สุดคือร้อยละ 87.39

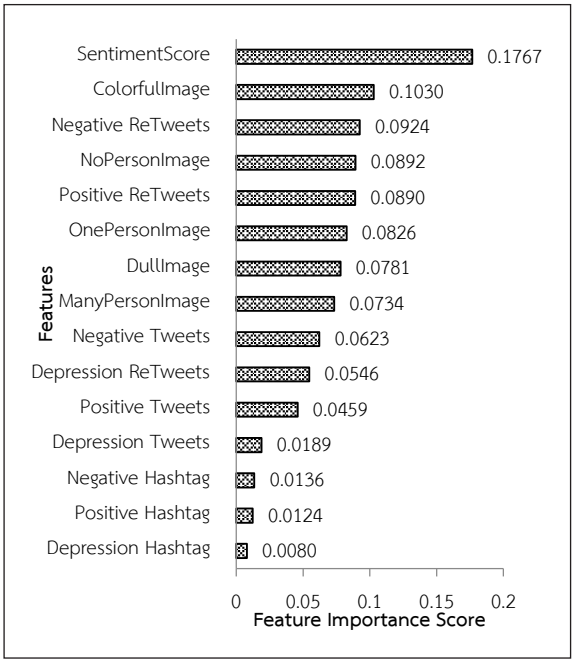
3.3 การหาคุณลักษณะที่สำคัญ (Feature Importance)

การหาคุณลักษณะที่สำคัญจากทวิตเตอร์สำหรับการสร้างแบบจำลองโดยพิจารณาจากค่าคะแนนความสำคัญ (Feature Importance Score) ซึ่งคำนวณได้จากผลรวมของค่าความสำคัญของคุณลักษณะบนต้นไม้แต่ละต้นจะคำนวณและหารด้วยจำนวนต้นไม้ทั้งหมด ดังสมการที่ (10)

$$RFfi_i = \frac{\sum_{j \in AllTrees} normfi_{ij}}{T} \quad (10)$$

โดยที่

$normfi_{ij}$ คือ ค่าคะแนนความสำคัญที่ทำให้เป็นมาตรฐาน (Normalized Feature Importance) สำหรับ โหนด i ในต้นไม้ j



รูปที่ 6 คุณลักษณะที่สำคัญสำหรับการสร้างแบบจำลอง

T คือ จำนวนต้นไม้ทั้งหมด

ค่าคะแนนความสำคัญ เป็นการลดลงของสิ่งเจือปนโหนดถ่วงน้ำหนัก (Impurity Weighted) โดยความน่าจะเป็นที่จะไปถึงโหนดนั้น ความน่าจะเป็นของโหนดคำนวณได้จากจำนวนตัวอย่างที่ไปถึงโหนด หารด้วยจำนวนตัวอย่างทั้งหมด ยิ่งค่าสูงคุณสมบัติยิ่งสำคัญ ดังรูปที่ 6

คุณลักษณะที่สำคัญที่สุดในการสร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์ ในงานวิจัยนี้ คือ ค่าเฉลี่ยคะแนนความรู้สึกของทวีต (Sentiment Score) รองลงมาคือ จำนวนรูปภาพสีสดใส (Colorful Image) จำนวนรีทวีตด้านลบ (Negative Retweet) และจำนวนรูปภาพที่ไม่มีคน (No Person Image)

ในขณะที่จำนวนแฮชแท็ก ประกอบด้วย จำนวนแฮชแท็กด้านบวก (Positive Hashtag) จำนวนแฮชแท็กด้านลบ (Negative Hashtag) และจำนวนแฮชแท็กที่แสดงถึงภาวะซึมเศร้า (Depression Hashtag) คือคุณลักษณะที่สำคัญน้อยที่สุดในการสร้างแบบจำลองในงานวิจัยนี้

4. อภิปรายผลและสรุป

จากผลการทดลองสามารถอภิปรายผลได้ดังนี้ ผลการเปรียบเทียบค่าประสิทธิภาพแบบจำลองพบว่า เทคนิคการสุ่มป่าไม้ให้ค่าประสิทธิภาพโดยรวมสูงที่สุด รองลงมาคือ เทคนิคต้นไม้การตัดสินใจ เพราะเทคนิคการสุ่มป่าไม้เป็นเทคนิคที่ใช้ได้ทั้งกับข้อมูลแบบมีโครงสร้าง (Structured Data) และข้อมูลที่ไม่มีโครงสร้าง (Unstructured Data) [19] อีกทั้งยังเป็นการพัฒนามาจากเทคนิคต้นไม้การตัดสินใจ สอดคล้องกับงานวิจัยของ Reece และคณะ [20] ที่สร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าด้วยเทคนิคการสุ่มป่าไม้ ซึ่งมีค่าประสิทธิภาพโดยรวมอยู่ที่ร้อยละ 77.20 โดยมีค่าน้อยกว่าเทคนิคที่พัฒนาขึ้นในงานวิจัยนี้ เนื่องจากความแตกต่างของข้อมูลนำเข้า (Input Data) และปริมาณของข้อมูลที่ใช้ในการสร้างแบบจำลอง ในขณะที่เทคนิคนาอีฟเบย์ (Naïve Bayes; NB) ให้ค่าประสิทธิภาพโดยรวมต่ำที่สุด ซึ่งเทคนิคนาอีฟเบย์นั้นเหมาะสำหรับการจำแนกชุดข้อมูลขนาดใหญ่ [21] ในงานวิจัยนี้ข้อมูลที่นำมาใช้ในการสร้างแบบจำลองอาจยังมีจำนวนข้อมูลไม่มากพอ

เมื่อเปรียบเทียบประสิทธิภาพของแบบจำลองจากการใช้ตัวแปรนำเข้าที่แตกต่างกันพบว่า คุณลักษณะที่เหมาะสมในการสร้างแบบจำลองในงานวิจัยนี้ คือ คุณลักษณะจากทวิตเตอร์ทั้งประเภทข้อความและประเภทรูปภาพร่วมกัน ซึ่งให้ค่าประสิทธิภาพโดยรวม (F-measure) สูงสุด คือ ร้อยละ 87.39 สอดคล้องกับค่าคะแนนความสำคัญของคุณลักษณะที่พบว่า คุณลักษณะที่สำคัญในการสร้างแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้ข้อมูลเครือข่ายสังคมออนไลน์อันดับแรกคือ ค่าเฉลี่ยคะแนนความรู้สึกของทวีต (Sentiment Score) ซึ่งเป็นคุณลักษณะที่อยู่ในคุณลักษณะจากทวิตเตอร์ประเภทข้อความ และรองลงมาคือ จำนวนรูปภาพสีสดใส ซึ่งเป็นคุณลักษณะจากทวิตเตอร์ประเภทรูปภาพ

อย่างไรก็ตามข้อมูลที่ใช้ในการทดลองนี้อยู่ในบริบทที่แตกต่างกับงานวิจัยอื่น ไม่ว่าจะเป็นคุณลักษณะของกลุ่มตัวอย่าง อาทิ เชื้อชาติ ภาษาที่ใช้ในการทวีต โดยใช้กระบวนการแปลภาษาจากภาษาไทยเป็นภาษาอังกฤษ ซึ่งอาจมีบางส่วนแปลได้คลาดเคลื่อนจากบริบทจริงของประโยคและนอกจากนั้นยังใช้คุณลักษณะที่สกัดมาใช้จากรูปภาพในทวิตเตอร์ ดังนั้นเทคนิคที่พัฒนาขึ้นมาจะเหมาะสมที่สุดกับข้อมูลคุณลักษณะชุดนี้

งานวิจัยนี้นำเสนอแนวคิดในการออกแบบและพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าโดยใช้เทคนิคการเรียนรู้ของเครื่อง ซึ่งสามารถนำแบบจำลอง และคุณลักษณะที่สำคัญในงานวิจัยนี้ไปใช้ในการวิจัยในอนาคตสำหรับการพัฒนาแบบจำลองการวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้าด้วยเทคนิคการเรียนรู้ของเครื่องแบบอื่น ๆ และใช้เป็นตัวแบบในการพัฒนาระบบวิเคราะห์ความเสี่ยงของการเกิดภาวะซึมเศร้า และใช้ในการป้องกันและลดความเสี่ยงการเกิดภาวะซึมเศร้าได้

กิตติกรรมประกาศ

งานวิจัยนี้ได้รับเงินอุดหนุนการวิจัยจากมหาวิทยาลัยเทคโนโลยีสุรนารี สำนักงานคณะกรรมการส่งเสริมวิทยาศาสตร์ วิจัยและนวัตกรรม และกองทุนส่งเสริมวิทยาศาสตร์ วิจัยและนวัตกรรม (รหัสโครงการ NRIIS 160345)

เอกสารอ้างอิง

[1] The Excellence Center for Depression Disorder, *Knowledge and essence about depression, World Health Day 2017*. Depression: Let's talk, 2017, pp 1–2.

[2] PUEY UNGPHAKORN Institute for Economic Research. (2021, May). *Mental health problems in Thailand during the Covid-19 crisis from the perspective of an economist.* [Online]. (in Thai). Available: https://www.pier.or.th/abridged/2021/08/

[3] Department of mental health, “Strategic Plan for the Department of Mental Health during the 12th National Economic and Social

Development Plan (2017–2021)," Mental Health Strategy Department of Mental Health, 2021, pp 19–20.

[4] Institute for population and social research Mahidol University, *Thai health report 2560.* Bangkok: Amarin Printing & Publishing Public Company Limited, 2017, pp 88–89 (in Thai).

[5] K. Phanichsiri and B. Tuntasood, "Social media addiction and attention deficit and hyperactivity symptoms in high school students in bangkok," *Journal of the Psychiatrist Association of Thailand*, vol. 61, no. 3, pp. 191–2004, 2016 (in Thai).

[6] M. Aldarwish and H. Ahmad, "Predicting depression levels using social media posts," in *IEEE 13th International Symposium on Autonomous Decentralized System (ISADS)*, Bangkok, Thailand, 2017, pp. 277–280.

[7] M. Choudhury, S. Counts and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proceedings of the 5th Annual ACM Web Science Conference (WebSci '13),* New York, USA: ACM, 2013, pp. 47–56.

[8] P. Vateekul and T. Koomsubha, "A study of sentiment analysis using deep learning techniques on Thai Twitter data," in *Proceeding of 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2016, pp 1–6.

[9] M. Park, C. Cha and M. Cha, "Depressive moods of users portrayed in Twitter," In *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD).* , SIGKDD, 2012, pp. 1-8.

[10] A.G. Reece and C.M. Danforth, "Instagram photos reveal predictive markers of depression," *EPJ Data Science*, vol. 6, no. 15, pp. 1–12. 2017.

[11] Department of mental health. (2018). *Patient Health Questionnaire-9 (9Q)*. [Online]. (in Thai). Available: https://www.dmh.go.th/test/download/files/2Q%209Q%208Q%20(1).pdf

[12] A. Esuli and F. Sebastiani, "SENTIWORDNET: A publicly available lexical resource for opinion mining," in *Proceedings of LREC-06, the 5th Conference on Language Resources*, 2006, pp. 417–422.

[13] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource forsentiment analysis and opinion mining," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2010, pp. 2200–2204.

[14] C. Musto, G. Semeraro and M. Polignano, "A comparison of Lexicon-based approaches for Sentiment Analysis of microblog posts," in *CEUR Workshop Proceedings*, 2014, 1314, pp. 59–68.

[15] V. A. Rao, K. Anuranjana, and R. Mamidi, "A SentiWordNet strategy for curriculum learning in sentiment analysis," *eprint arXiv*, 2005.04749, 2020.

[16] S. Tipprasert. (2022). *Depression dataset*. [Online]. (in Thai). Available: https://www.twothai.com/dep/.

[17] P. Viola and M. Jones, (2001). "Rapid object detection using a Boosted cascade of simple features," in *Proceedings of the IEEE conference on computer vision and pattern Recognition*, 2001, pp. 1–9.

[18] L. Wilms and D. Oberfeld, "Color and emotion: effects of hue, saturation, and brightness," *Psychological Research*, vol. 82, pp. 896–914, 2018.

[19] P. Sanguansat. *Artificial Intelligence with Machine Learning, Digital Technology*. 1st ed. Nonthaburi: IDC Premier, 2019 (in Thai).

[20] A. G. Reece, A. J. Reagan, K. L. M. Lix, P. S. Dodds, C.M. Danforth and E.J. Langer, "Forecasting the onset and course of mental illness with Twitter data," *Scientific Reports* 7, pp. 13006, 2017.

[21] M. F. Kabir, C. M. Rahman, A. Hossain and K. Dahal, "Enhanced classification accuracy on naive bayes data mining models," *International Journal of Computer Applications*, vol. 28, no. 3, pp 9–16. 2011.